DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales and Javier Ortega-Garcia
Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, Spain
{ruben.tolosana, ruben.vera, julian.fierrez, aythami.morales, javier.ortega}@uam.es

***Abstract*—The free access to large-scale public databases, together with the fast progress of deep learning techniques, in particular Generative Adversarial Networks, have led to the generation of very realistic fake contents with its corresponding implications towards society in this era of fake news.**

This survey provides a thorough review of techniques for manipulating face images including DeepFake methods, and methods to detect such manipulations. In particular, four types of facial manipulation are reviewed: *i)* entire face synthesis, *ii)* face identity swap (DeepFakes), *iii)* facial attributes manipulation, and *iv)* facial expression manipulation. For each manipulation type, we provide details regarding manipulation techniques, existing public databases, and key benchmarks for technology evaluation of fake detection methods, including a summary of results from those evaluations. Among the different databases available and discussed in the survey, FaceForensics++ is for example one of the most widely used for detecting both face identity swap and facial expression manipulations, with results in the literature in the range of 90-100% of manipulation detection accuracy.

In addition to the survey information, we also discuss trends and provide an outlook of the ongoing work in this field, e.g., the recently announced DeepFake Detection Challenge (DFDC).

***Index Terms*—Fake news, DeepFakes, Face manipulation, Databases, Benchmark, Face recognition**

I. INTRODUCTION

FAKE images and videos including facial information generated by digital manipulation, in particular with DeepFake methods [1], have become a great public concern recently [2], [3]. The very popular term "DeepFake" is referred to a deep learning based technique able to create fake images/videos by swapping the face of a person in an image or video by the face of another person. This term was originated after a Reddit user named "deepfakes" claimed in late 2017 to have developed a machine learning algorithm that helped him to transpose celebrity faces into porn videos [4]. Apart from fake pornography, some of the more harmful usages of such fake content include fake news, hoaxes, and financial fraud. As a result, the area of research traditionally dedicated to general image and video fake detection [5], is being invigorated and is now dedicating growing efforts for detecting facial manipulation in image and video [6], [7]. These renewed efforts in fake face detection are built around past research in biometric anti-spoofing [8]–[10] and modern data-driven deep learning [11]. The growing interest in fake face detection is demonstrated through the increasing number of workshops in top conferences [12]–[14] and competitions such as the recent MFC2018[1] and DFDC[2] launched by NIST and Facebook, respectively.

Traditionally, the number and realism of facial manipulations have been limited by the lack of sophisticated editing tools, the domain expertise required, and the complex and time-consuming process involved. For example, an early work in this topic [15] was able to modify the lip motion of a person speaking using a different audio track, by making connections between the sounds of the audio track and the shape of the subject's face. Nowadays, it is becoming increasingly easy to automatically synthesise non-existent faces or manipulate a real face of one person in an image/video, thanks to: *i)* the accessibility to large public data, and *ii)* the evolution of deep learning techniques that eliminate manual editing steps.

An evolution of the previous technology was presented in [16], generating high-quality videos of a person (Obama in this case) changing what he is really saying in a target video[3]. Nowadays, the most realistic manipulation techniques that have superseded the methods in [16] make use of Generative Adversarial Networks (GANs) with Convolutional Neural Networks (CNNs). As a result, open software and mobile application such as ZAO[4] and FaceApp[5] have been released opening the door to anyone to create fake images and videos. In response to those increasingly sophisticated and realistic manipulated contents, large efforts are being carried out nowadays by the research community to create improved methods for face manipulation detection.

This survey provides a thorough review of four types of facial manipulation: *i)* entire face synthesis, *ii)* face identity swap, *iii)* facial attributes manipulation, and *iv)* facial expression manipulation. For each manipulation type, we provide details regarding manipulation techniques, existing public databases, and key benchmarks for technology evaluation of fake detection methods, including a summary of results from those evaluations.

The remainder of the article is organised as follows. We first provide in Sec. II a general description of different types of facial manipulation. Then, from Sec. III to Sec. VI we describe the key elements of each type of facial manipulation including public databases for research, detection methods,

[1] https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018
[2] https://deepfakedetectionchallenge.ai/
[3] https://www.youtube.com/watch?v=o2DDU4g0PRo
[4] https://apps.apple.com/cn/app/id1465199127
[5] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341

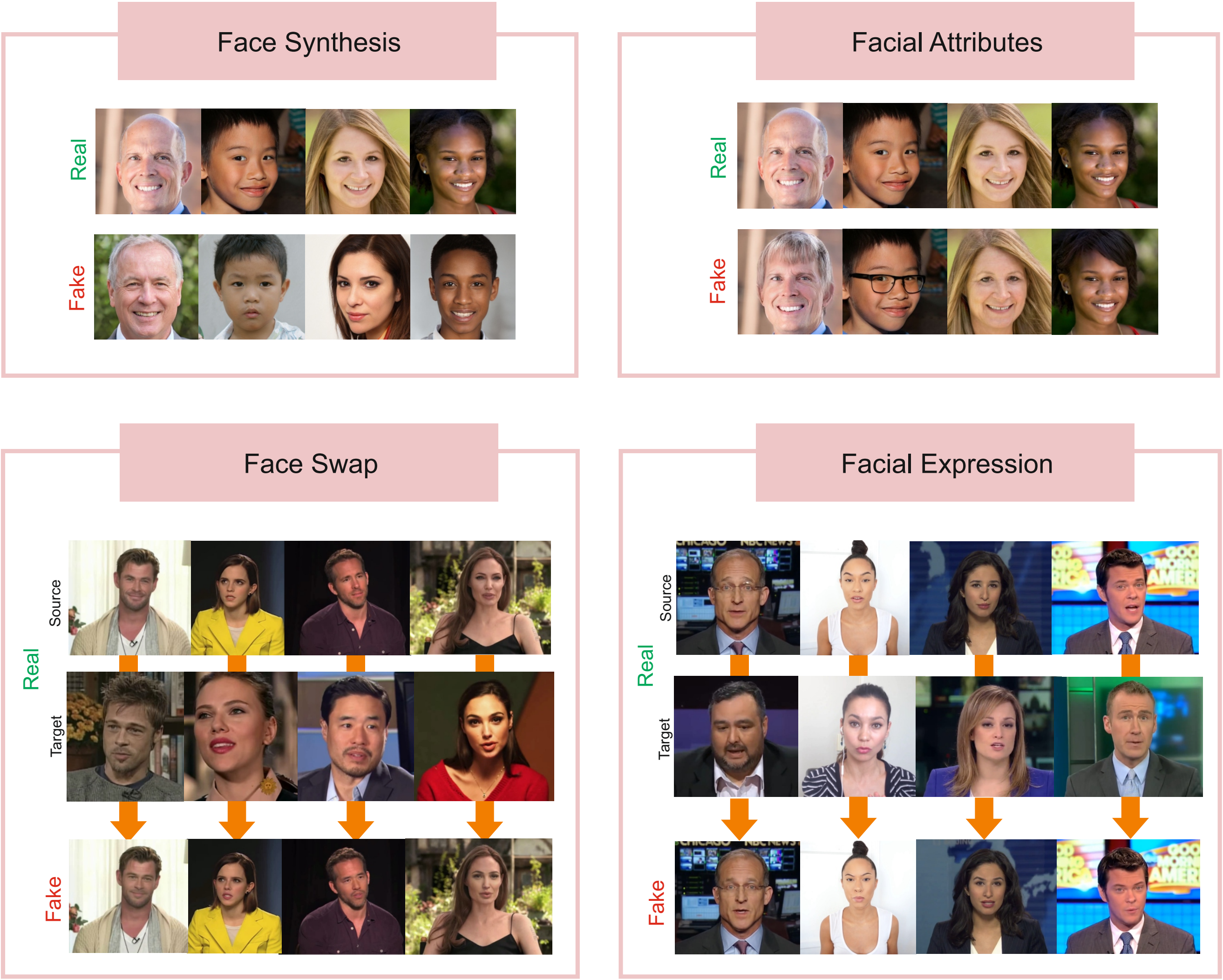

Fig. 1. Real and fake examples of each facial manipulation group. For *Face Synthesis*, real images are extracted from http://www.whichfaceisreal.com/ and fake images from https://thispersondoesnotexist.com. For *Face Swap*, face images are extracted from Celeb-DF database [17]. For *Facial Attributes*, real images are extracted from http://www.whichfaceisreal.com/ and fake images are generated using FaceApp. Finally, for *Facial Expression*, images are extracted from FaceForensics++ [6].

and benchmark results. Finally, we provide in Sec. VII our concluding remarks, highlighting future research lines.

II. Types of Facial Manipulations

Facial manipulations can be categorised in four different groups regarding the level of manipulation [7]. Fig. 1 graphically summarises each facial manipulation group. We now describe each of them, from higher to lower level of manipulation:

- **Face synthesis:** this manipulation creates entire non-existent faces, usually through powerful Generative Adversarial Networks (GANs) [18], e.g., through the recent StyleGAN approach proposed in [19]. These techniques achieve astonishing results, generating high-quality facial images with a high level of realism. Fig. 1 shows some examples for entire face synthesis extracted using StyleGAN[6].
- **Face swap:** this manipulation consists of replacing the face of one person with the face of another person. Two different approaches are usually considered: *i)* classical computer graphics-based techniques such as FaceSwap[7], and *ii)* novel deep learning techniques known as DeepFakes[8], e.g., the recent ZAO mobile application. Popular and very realistic videos of this type of manipulation can be seen on Youtube[9].
- **Facial attributes:** this manipulation consists of modifying some attributes of the face such as the colour of the hair or the skin, the gender, the age, adding glasses, etc [20]. This manipulation process is usually carried out through GANs such as the StarGAN approach proposed in [21]. One example of this type of manipulation is the popular FaceApp mobile application.

[6] https://thispersondoesnotexist.com

[7] https://github.com/MarekKowalski/FaceSwap

[8] https://github.com/deepfakes/faceswap

[9] https://www.youtube.com/watch?v=UlvoEW7l5rs

- **Facial expression:** this manipulation consists of modifying the facial expression of the person, e.g., transferring the facial expression of one person to another person. One of the most popular techniques is Face2Face [22], working on real time. Recent approaches have shown its potential, generating high-quality videos of a person (Obama) changing what he is really saying in a target video [16].

III. Face Synthesis

A. Manipulation Techniques and Public Databases

Table I summarises the main publicly available databases for research focused on the entire face synthesis. It is important to highlight that none of them provide real face image examples. Researchers focused on this type of manipulation usually consider real faces from popular public databases such as CelebA [23], FFHQ [19], CASIA-WebFace [24], and VGGFace2 [25], among others to train their systems.

Regarding fake images, four different databases are noteworthy. In [19], Karras *et al.* released a set of 100,000 synthetic face images, named 100K-Generated-Images[10]. This database was generated using their proposed StyleGAN architecture [19], which was trained using the FFHQ dataset [19]. StyleGAN is an improved version of their previous popular approach ProGAN [26], which introduced a new training methodology based on improving both generator and discriminator progressively. StyleGAN proposes an alternative generator architecture that leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis.

Another public database is 100K-Faces [27]. This database contains 100,000 synthetic images generated using StyleGAN as well. In this database, contrary to the 100K-Generated-Images database, the StyleGAN network was trained using around 29,000 photos from 69 different models, considering face images from a more controlled scenario (e.g., with a flat background). Thus, no strange artifacts created by the StyleGAN are included in the background of the images.

Recently, Stehouwer *et al.* introduced in [7] a new database named Diverse Fake Face Dataset (DFFD), which is not publicly available up to date. Regarding the entire face synthesis manipulation, the authors created 100,000 and 200,000 fake images through the pre-trained ProGAN [26] and StyleGAN [19]) models, respectively.

Finally, Neves *et al.* presented in [11] the Face Synthetic Removal database (FSRemovalDB). This database comprises a total 150,000 synthetic face images originally created through StyleGAN. Contrary to the other databases, in this database the GAN "fingerprints" produced by the StyleGAN were removed from the original synthetic fake images through the use of autoencoders, while keeping the visual quality of the resulting images. Therefore, this database presents a higher level of manipulation for the detection systems.

TABLE I
Face synthesis: Publicly available databases.

Database	Real Images	Fake Images
100K-Generated-Images (2019) [19]	-	100,000 (StyleGAN)
100K-Faces (2019) [27]	-	100,000 (StyleGAN)
DFFD (2019) [7]	-	100,000 (StyleGAN) 200,000 (ProGAN)
FSRemovalDB (2019) [11]	-	150,000 (StyleGAN)

B. Manipulation Detection

Different studies have recently evaluated the difficulty of detecting manipulations based on the entire face synthesis. Table II depicts a comparison of the most relevant approaches in this area. For each study, we include information related to the features, classifiers, best performance, and databases considered. We highlight in **bold** the best results achieved for each public database. It is important to remark that in some cases, different evaluation metrics are considered (e.g., Area Under the Curve (AUC) or Equal Error Rate (EER)), which makes it difficult to perform a fair comparison among the studies.

In [28], the authors analysed the GAN pipeline in order to detect different artifacts between real and fake images. They proposed a detection system based on colour features and a linear Support Vector Machine (SVM) for the final classification. The proposed approach achieved a final 70.0% of AUC for the best performance when evaluating with the NIST MFC2018 dataset [33].

Later on, Yu *et al.* analysed in [29] the existence and uniqueness of GAN fingerprints in order to detect fake images. In particular, they proposed a learning-based formulation based on an attribution network architecture to map an input image to its corresponding fingerprint image. Therefore, they learned a model fingerprint for each source (each GAN instance plus the real world), such that the correlation index between one image fingerprint and each model fingerprint serves as softmax logit for classification. Their proposed approach was tested using real faces from CelebA database [23] and synthetic faces created through different GAN approaches (ProGAN [26], SNGAN [34], CramerGAN [35], and MMDGAN [36]), achieving a final 99.5% accuracy for the best performance. However, this approach seemed not to be very robust against simple image perturbation attacks such as noise, blur, cropping or compression, unless the models were re-trained again.

In [30], Wang *et al.* conjectured that monitoring neuron behavior could also serve as an asset in detecting fake faces since layer-by-layer neuron activation patterns may capture more subtle features that are important for the facial manipulation detection system. Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (i.e., VGG-Face [37], OpenFace [38], and FaceNet [39]), and then trained a SVM for the final classification. The authors tested their proposed approach using real faces from CelebA-HQ [26] and FFHQ [19] databases and synthetic faces created through

[10] https://github.com/NVlabs/stylegan

TABLE II
FACE SYNTHESIS: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK.
AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
McCloskey and Albright (2018) [28]	Colour-related	SVM	AUC = 70.0%	NIST MFC2018
Yu *et al.* (2019) [29]	GAN-related	CNN	Acc. = 99.5%	Own (ProGAN, SNGAN, CramerGAN, MMDGAN)
Wang *et al.* (2019) [30]	CNN Neuron Behavior	SVM	Acc. = 84.7%	Own (InterFaceGAN, StyleGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 100%** **EER = 0.1%**	**DFFD (ProGAN, StyleGAN)**
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	*EER = 7.2%*	*100K-Faces (StyleGAN)*
Neves *et al.* (2019) [11]	Image-related	CNN	**EER = 0.8%**	**100K-Faces (StyleGAN)**
			EER = 20.6%	**FSRemovalDB (StyleGAN)**
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN)

InterFaceGAN [40] and StyleGAN [19], achieving for the best performance a final 84.7% accuracy using the FaceNet model.

More recently, Stehouwer *et al.* carried out in [7] a complete analysis of different types of facial manipulations. They proposed to use attention mechanisms to process and improve the feature maps of CNN models. For the entire face synthesis manipulation, the authors achieved a final 100% AUC and around 0.1% Equal Error Rate (EER) considering real faces from CelebA [23], FFHQ [19], and FaceForensics++ [6] databases and fake images created through ProGAN [26] and StyleGAN [19] approaches. The impressive results achieved show the importance of novel attention mechanisms [41].

Nataraj *et al.* proposed in [31] a detection system inspired by steganalysis and natural image statistics. In particular, their method was based on a combination of pixel co-occurrence matrices and CNNs. Their proposed approach was initially tested through a database of various objects and scenes created through CycleGAN [42]. Besides, the authors performed an interesting analysis to see the robustness of the proposed approach against fake images created through different GAN architectures (CycleGAN vs. StarGAN), with good generalisation results. This detection approach was implemented later on in [11] considering images from the 100K-Faces database, achieving an EER of 7.2% for the best performance. This result is remarked in *italics* to indicate that it was not provided in the original paper.

In [11], Neves *et al.* performed a thorough experimental assessment of this type of facial manipulation considering state-of-the-art detection systems and different experimental conditions, i.e., controlled and in-the-wild scenarios. Two different fake databases were considered: *i)* a database comprised of 150,000 unique faces collected online[11], and *ii)* 100K-faces. In controlled scenarios, they achieved similar results as the best previous studies (EER = 0.8%). However, in more challenging scenarios in which images (real and fake) come from different sources (mismatch of datasets), a high degradation of the performance is observed. In addition, they also proposed a novel approach to remove the GAN "fingerprint" information from fake images so as to spoof state-of-the-art detection systems achieving EERs above 20%. The results achieved highlight that more efforts need to be done in order to develop more robust detection systems against unseen conditions.

Related to the unseen conditions just commented, Marra *et al.* performed in [32] an interesting study in order to detect unseen types of fake generated data. Concretely, they proposed a multi-task incremental learning detection method in order to detect and classify new types of GAN generated images, without worsening the performance on the previous ones. Two different solutions regarding the position of the classifier were proposed based on the successful algorithm iCaRL for incremental learning [43]: *i)* Multi-Task MultiClassifier (MT-MC), and *ii)* Multi-Task Single Classifier (MT-SC). Regarding the experimental framework, five different GAN approaches were considered in the study, CycleGAN [42], ProGAN [26], Glow [44], StarGAN [21], and StyleGAN [19]. Their proposed detection approach, based on the XceptionNet model, achieved promising results being able to correctly detect new GANs generated images presented to the network.

Finally, we also include for completeness some important references to other recent studies focused on the detection of general GAN-based image manipulations, not facial ones. In particular, we refer the reader to [45], [46].

IV. FACE SWAP

A. Manipulation Techniques and Public Databases

This is one of the most popular face manipulation techniques nowadays. Unlike the entire face synthesis manipulation, where just images are usually considered, in face swap the aim is usually to detect whether a video is real or fake. Table III summarises current public databases focused on this type of face manipulation. As can be seen, both real and fake videos are usually included in the databases.

One of the first public databases was UADFV [47]. This database comprises 49 real videos from Youtube, which were used to create 49 fake videos through the FakeApp mobile

[11] https://thispersondoesnotexist.com

TABLE III
FACE SWAP: PUBLICLY AVAILABLE DATABASES.

Database	Real Videos	Fake Videos
UADFV (2018) [47]	49 (Youtube)	49 (FakeApp)
DeepfakeTIMIT (2018) [1]	-	620 (faceswap-GAN)
FaceForensics++ (2019) [6]	1000 (Youtube)	1000 (FaceSwap) 1000 (DeepFake)
DeepFakeDetection (2019) [50]	363 (Actors)	3068 (DeepFake)
Celeb-DF (2019) [17]	408 (Youtube)	795 (DeepFake)
DFDC Preview (2019) [51]	1131 (Actors)	4119 (Unknown)

application. Each video represents one individual, with a typical resolution of 294×500 pixels, and 11.14 seconds on average.

Korshunov and Marcel introduced in [1] a new database named DeepfakeTIMIT. This database comprises 620 fake videos of 32 subjects from the VidTIMIT database [48]. Fake videos were created using the public GAN-based face-swapping algorithm[12]. In that approach, the generative network is adopted from CycleGAN [42], using the weights of FaceNet [39]. The method Multi-Task Cascaded Convolution Networks is used for more stable detections and reliable face alignment [49]. Besides, the Kalman filter is also considered to smooth the bounding box positions over frames and eliminate jitter on the swapped face. Regarding the scenarios considered in DeepfakeTIMIT, two different qualities are considered: *i)* low quality (LQ) with images of 64×64 pixels, and *ii)* high quality (HQ) with images of 128×128 pixels. Additionally, different blending techniques were applied to the fake videos regarding the quality level.

One of the most popular databases in this type of facial manipulation is FaceForensics++ [6]. This database was introduced in 2019 as an extension of the original FaceForensics database [61], which was focused only on facial expression manipulation. FaceForensics++ contains 1000 real videos extracted from Youtube. Regarding the face swapping/identity swap fake videos, they were generated using both computer graphics and DeepFake approaches (i.e., learning approach). For the computer graphics approach, the authors considered the publicly available FaceSwap algorithm[13] whereas for the DeepFake approach, fake videos were created through the DeepFake FaceSwap GitHub implementation[14]. The FaceSwap approach consists of face alignment, Gauss Newton optimization and image blending to swap the face of the source person to the target person. The DeepFake approach, as indicated in [6], is based on two autoencoders with a shared encoder that are trained to reconstruct training images of the source and the target face, respectively. A face detector is used to crop and to align the images. To create a fake image, the trained encoder and decoder of the source face are applied to the target face. The autoencoder output is then blended with the rest of the image using Poisson image editing [62]. Regarding the figures of the FaceForensics++ database, 1000 fake videos were generated for each approach. Later on, a new dataset named DeepFakeDetection was included inside the FaceForensics++ framework with the support of Google [50]. This dataset comprises 363 real videos from 28 paid actors in 16 different scenes. Additionally, 3068 fake videos are included in the dataset based on DeepFake FaceSwap GitHub implementation. It is important to remark that for both FaceForensics++ and DeepFakeDetection databases different levels of video quality are considered, in particular: *i)* RAW (original quality), *ii)* HQ (constant rate quantization parameter equal to 23), and *iii)* LQ (constant rate quantization parameter equal to 40). This aspect simulates the video processing techniques usually applied in social networks.

Recently, Li *et al.* have presented in [17] a new database named Celeb-DF. This database aims to provide fake videos of better visual qualities, similar to the popular videos that are available online[15], as previous databases exhibit low visual quality with many visible artifacts. Celeb-DF consists of 408 real videos extracted from Youtube, and 795 fake videos, which were created through a refined version of a public DeepFake generation algorithm, improving aspects such as the low resolution of the synthesised faces and colour inconsistencies.

Finally, Facebook in collaboration with other companies and academic institutions such as Microsoft, Amazon, and the MIT have recently launched a new challenge named the Deepfake Detection Challenge (DFDC) [51]. They first released a preview dataset consisting of 1131 real videos from 66 paid actors, and 4119 fake videos. Fake videos were generated using two different unknown approaches. The complete DFDC dataset (over 470 GB) was released on 11th of December with the beginning of the competition. So far today, over 500 competitors are currently participating in the challenge[16].

B. Manipulation Detection

The development of novel methods to detect face swap manipulations is continuously evolving, being one of the most studied facial manipulation groups. Table IV provides a comparison of the most relevant detection approaches in this area. For each study we include information related to the features, classifiers, best performance, and databases for research. We highlight **in bold** the best results achieved for each public database. It is important to remark that in some cases, different evaluation metrics are considered (e.g., AUC and EER), which makes it difficult to perform a fair comparison among the studies. Finally, the results highlighted in *italics* indicate the generalisation capacity of the detection systems against different unseen databases, i.e., those databases were not considered during training. These results have been extracted from [17] and were not included in the original publications.

One of the first detection approaches proposed in the literature was [52]. In that study, Zhou *et al.* proposed a two-

[12] https://github.com/shaoanlu/faceswap-GAN
[13] https://github.com/MarekKowalski/FaceSwap
[14] https://github.com/deepfakes/faceswap
[15] https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g
[16] https://www.kaggle.com/c/deepfake-detection-challenge

TABLE IV
Face swap: Comparison of different state-of-the-art detection approaches. The best results achieved for each public database are remarked in **bold**. Results in *italics* indicate that they were not provided in the original work.
FF++ = FaceForensics++, AUC = Area Under the Curve, Acc. = Accuracy, EER = Equal Error Rate.

Study	Features	Classifiers	Best Performance	Databases
Zhou *et al.* (2018) [52]	Image-related Steganalysis	CNN SVM	*AUC = 85.1%*	*UADFV*
			AUC = 83.5%	*DeepfakeTIMIT (LQ)*
			AUC = 73.5%	*DeepfakeTIMIT (HQ)*
			AUC = 70.1%	*FF++ / DFD*
			AUC = 55.7%	***Celeb-DF***
Afchar *et al.* (2018) [53]	Mesoscopic Level	CNN	Acc. = 98.4%	Own
			AUC = 84.3%	*UADFV*
			AUC = 87.8%	*DeepfakeTIMIT (LQ)*
			AUC = 62.7%	*DeepfakeTIMIT (HQ)*
			Acc. ≃ 90.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 94.0%	FF++ (DeepFake, HQ)
			Acc. ≃ 98.0%	FF++ (DeepFake, RAW)
			Acc. ≃ 83.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 93.0%	FF++ (FaceSwap, HQ)
			Acc. ≃ 96.0%	FF++ (FaceSwap, RAW)
			AUC = 53.6%	*Celeb-DF*
Korshunov and Marcel (2018) [1]	Lip Image - Audio Speech Image-related	PCA+RNN PCA+LDA, SVM	**EER = 3.3%** **EER = 8.9%**	**DeepfakeTIMIT (LQ)** **DeepfakeTIMIT (HQ)**
Güera and Delp (2018) [54]	Image + Temporal Information	CNN + RNN	Acc. = 97.1%	Own
Yang *et al.* (2019) [55]	Head Pose Estimation	SVM	AUC = 89.0%	UADFV
			AUC = 55.1%	*DeepfakeTIMIT (LQ)*
			AUC = 53.2%	*DeepfakeTIMIT (HQ)*
			AUC = 47.3%	*FF++ / DFD*
			AUC = 54.8%	*Celeb-DF*
Li *et al.* (2019) [56]	Face Warping Artifacts	CNN	**AUC = 97.4%**	**UADFV**
			AUC = 99.9%	**DeepfakeTIMIT (LQ)**
			AUC = 93.2%	**DeepfakeTIMIT (HQ)**
			AUC = 79.2%	*FF++ / DFD*
			AUC = 53.8%	*Celeb-DF*
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 94.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 98.0%	**FF++ (DeepFake, HQ)**
			Acc. ≃ 100.0%	**FF++ (DeepFake, RAW)**
			Acc. ≃ 93.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 97.0%	**FF++ (FaceSwap, HQ)**
			Acc. ≃ 99.0%	**FF++ (FaceSwap, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 70.2%	*UADFV*
			AUC = 77.0%	*DeepfakeTIMIT (LQ)*
			AUC = 77.3%	*DeepfakeTIMIT (HQ)*
			AUC = 78.0%	*FF++ / DFD*
			AUC = 48.8%	*Celeb-DF*
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	*AUC = 65.8%*	*UADFV*
			AUC = 62.2%	*DeepfakeTIMIT (LQ)*
			AUC = 55.3%	*DeepfakeTIMIT (HQ)*
			AUC = 76.3%	*FF++ / DFD*
			EER = 15.1%	FF++ (FaceSwap, HQ)

Stehouwer *et al.* (2019)
[7]
	Image-related	CNN + Attention Mechanism	AUC = 99.4%
			EER = 3.1%
				DFFD

Dolhansky *et al.* (2019) [51]	Image-related	CNN	**Precision = 93.0%** **Recall = 8.4%**	**DFDC Preview**
Agarwal and Farid (2019) [59]	Facial Expressions and Pose	SVM	AUC = 96.3%	Own (FaceSwap, HQ)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**AUC = 96.9%** **AUC = 96.3%**	**FF++ (DeepFake, LQ)** **FF++ (FaceSwap, LQ)**

stream network for face manipulation detection. In particular, the authors considered a fusion of two streams: *i)* a face classification stream based on the CNN GoogLeNet [63] to detect whether a face image is fake or not, and *ii)* a path triplet stream that is trained using steganalysis features of images patches with a triplet loss, and a SVM for the classification. The initial system was trained to detect facial expression manipulations. Nevertheless, Li *et al.* evaluated in [17] the generalisation capacity of the pre-trained model (using SwapMe app) to detect face swapping/identity swap face manipulations, resulting to be one the most robust approaches against the recent Celeb-DF database [17].

Later on, Afchar *et al.* proposed in [53] two different networks composed of few layers in order to focus on the mesoscopic properties of the images: *i)* a CNN network comprised of 4 convolutional layers followed by a fully-connected layer (Meso-4), and *ii)* a modification of Meso-4 consisted of a variant of the Inception module introduced in [63], named MesoInception-4. Their proposed approach was originally tested against DeepFakes using a private database, achieving a 98.4% of accuracy for the best performance. That pre-trained detection model was tested against unseen databases in [17], proving to be a robust approach in some cases such as with the FaceForensics++ database.

In [1], Korshunov and Marcel first showed that state-of-the-art face recognition systems such as VGG [37] and FaceNet [39] are vulnerable to DeepFake videos from the DeepfakeTIMIT database. Then, they evaluated how challenging is to detect fake videos using baseline approaches based on inconsistencies between lip movements and audio speech, as well as, several variations of image-based systems. For the first case, they considered Mel-Frequency Cepstral Coefficients (MFCCs) as audio features and distances between mouth landmarks as visual features. Principal Component Analysis (PCA) was then used to reduce the dimensionality of the blocks of features, and finally Recurrent Neural Networks (RNNs) based on Long Short-Term Memory (LSTM) to detect real of fake videos (based on [64]). For the second case, they evaluated detection approaches based on: *i)* raw faces as features, and *ii)* image quality measures (IQM) [65]. In particular, they used a set of 129 features related to measures like signal to noise ratio, specularity, blurriness, etc. PCA with Linear Discriminant Analysis (LDA), or SVM were considered for the final classification. Their proposed detection approach based on IQM+SVM provided the best results, with a final 3.3% and 8.9% of EER for the LQ and HQ scenarios of the DeepfakeTIMIT database, respectively.

Güera and Delp proposed in [54] a temporal-aware pipeline to automatically detect fake videos. They considered a combination of CNNs and RNNs. For the CNN, the authors used InceptionV3 [66] pre-trained using ImageNet database [67]. For the RNN system, they considered a LSTM model composed of one hidden layer with 2048 memory blocks. Finally, two fully-connected layers were included, providing the probabilities of the frame sequence being either real or fake. Their proposed approach was evaluated using a proprietary database with a final 97.1% accuracy.

Yang *et al.* observes in [55] that current DeepFakes are created by splicing synthesised face regions into the original image, and in doing so, introducing errors that can be revealed when 3D head poses are estimated from the face images. Thus, they performed an study based on the differences between head poses estimated using a full set of facial landmarks (68 extracted from DLib [68]) and those in the central face regions to differentiate DeepFakes from real videos. Once these features are extracted and normalised (mean and standard deviation), a SVM is considered for the final classification. Their proposed approach was originally evaluated with the UADFV database, achieving a final AUC of 89.0%. However, this pre-trained model (using UADFV database) seems not to generalise very well to other databases as depicted in Table IV.

The same authors proposed in [56] another approach based on the detection of face warping artifacts. The motivation is that current DeepFake generation algorithms can only create images of limited resolution, which need to be further warped to match the original faces in the source video. Such transforms leave distinctive artifacts in the resulting DeepFake videos. Thus, the authors proposed a detection system based on CNNs in order to detect the presence of such artifacts from the detected face regions and the surrounding areas. Four different CNN models were trained from scratch, i.e., VGG16 [69], ResNet50, ResNet101, and ResNet152 [70]. Their proposed detection approach was tested using the UADFV and DeepfakeTIMIT databases, outperforming the state of the art for those databases.

An exhaustive analysis of face swap was carried out by Rössler *et al.* in [6]. Four different detection systems were evaluated using the FaceForensics++ database: *i)* a CNN-based system trained through handcrafted steganalysis features [71], *ii)* a CNN-based system whose convolution layers are specifically designed to suppress the high-level content of the image [72], *iii)* a CNN-based system with a global pooling layer that computes four statistics (mean, variance, maximum, and minimum) [73], *iv)* the CNN MesoInception-4 detection system described in [53], and finally *v)* the CNN-based system XceptionNet [74] pre-trained using ImageNet database [67] and re-trained for the face manipulation task. In general, the detection system based on XceptionNet architecture provided the best results in both types of manipulation methods, DeepFakes and FaceSwap. In addition, the detection systems were evaluated considering different video quality levels in order to simulate the video processing of many social networks. In this real scenario, the accuracy of all detection systems decreased when lowering the video quality, remarking how challenging is this task in real scenarios.

In [57], Matern *et al.* first revised current facial manipulation methods highlighting several characteristic artifacts that appear from their corresponding pipelines. Besides, they proposed a detection method based on relatively simple visual artifacts such as eye colour, missing reflections, and missing details in the eye and teeth areas. Two different classifiers are considered in this analysis: *i)* a logistic regression model, and *ii)* a Multilayer Perceptron (MLP) [75]. Their proposed approach was tested using a private database, achieving a final 85.1% of AUC for the MLP system.

Another interesting approach was proposed by Nguyen *et* *al.* in [58]. In that study, the authors proposed a CNN system that uses multi-task learning to simultaneously detect fake videos and locate the manipulated regions. They considered a detection system based on an autoencoder. Concretely, they proposed to use a Y-shaped decoder in order to share valuable information between the classification, segmentation, and reconstruction tasks, improving the overall performance by reducing the loss. Their proposed approach was evaluated with the FaceSwap manipulation method for the FaceForensics++ database [61], achieving a best performance of 15.07% EER, a performane not as good as for other approaches. In addition, this model seems not to generalise very well for other databases, with results below 80% AUC.

Stehouwer *et al.* performed in [7] a thorough analysis of different face manipulations. They proposed a detection system based on CNN and attention mechanisms to process and improve the feature maps of the classifier model. Their proposed attention map can be implemented easily and inserted into existing backbone networks, through the inclusion of a single convolution layer, its associated loss functions, and masking the subsequent high-dimensional features. Their proposed detection approach was tested with the DFFD database (based in this type of manipulation on a combination of previous databases such as FaceForensics++ and a collection of internet data). In particular, for face swap detection, their proposed approach achieved an AUC of 99.43% and EER of 3.1%. Despite of the fact that it is difficult to provide a fair comparison among studies as different experimental protocols are considered, it is clear that their detection approach provides state-of-the-art results.

As a result of the high popularity and importance of the topic, Facebook in collaboration with other companies and academic institutions have recently launched the DeepFake Detection Challenge competition. In [51], in addition to the description of the database, the authors have provided baseline results using three simple detection systems: *i)* a small CNN model composed of 6 convolution layers and 1 fully-connected layer to detect low-level image manipulations, *ii)* a XceptionNet model trained using only face images, and *iii)* a XceptionNet model trained using the full image. The detection system based on XceptionNet with only the face image provides the best results with 93.0% precision and 8.4% recall.

Agarwal and Farid proposed in [59] a detection technique based on facial expressions and head movements. For the feature extraction, the OpenFace2 toolkit was considered [76], obtaining an intensity and occurrence for 18 different facial action units related to movements of facial muscles such as cheek raiser, nose wrinkler, mouth stretch, etc. Additionally, four features related to head movements were considered. As a result, each 10-second video clip is reduced to a feature vector of dimension 190 using the Pearson correlation to measure the linearity between features. Finally, the authors considered a SVM for the final classification. Regarding the experimental framework, the authors build their own database based on videos downloaded from YouTube of persons of interest talking in a formal setting, for example, weekly address, news interview, and public speech. In most videos the person is primarily facing towards the camera. Regarding the DeepFake videos, the authors trained one GAN per person based on faceswap-GAN[17]. Their proposed approach achieved a final AUC of 96.3% for the best performance, being robust against new contexts and manipulation techniques.

Finally, Sabir *et al.* [60] proposed a method to detect fake videos based on using the temporal information present in the stream. The intuition behind this model is to exploit temporal discrepancies across frames. Thus, they considered a recurrent convolutional network similar to [54], trained in this study end-to-end instead of using a pre-trained model. Their proposed detection approach was tested through FaceForensics++ database, achieving AUC results of 96.9% and 96.3% for the DeepFake and FaceSwap methods, respectively. Only the low-quality videos were considered in the analysis.

V. Facial Attributes

A. Manipulation Techniques and Public Databases

Despite the success of GAN-based frameworks for general image translations and manipulations [16], [21], [42], [82]–[85], and in particular for face attribute manipulations [21], [86]–[92], few databases are publicly available to the best of our knowledge. The main reason is that the code of most GAN approaches are publicly available, so researchers can easily generate their own fake databases as they like. Therefore, this section aims to highlight the latest GAN approaches in the field, from older to more recent ones, providing also the link to their corresponding GitHub codes.

In [86], the authors introduced the Invertible Conditional GANs (IcGANs)[18] for complex image editing as the union of an encoder used jointly with a conditional GAN (cGAN) [93]. This approach provides accurate results in terms of editing attributes. However, it seriously changes the face identity of the person.

Lample *et al.* proposed in [89] an encoder-decoder architecture that is trained to reconstruct images by disentangling the salient information of the image and the attribute values directly in the latent space[19]. However, as it happens with the IcGAN approach, the generated images may lack some details or present unexpected distortions.

An enhanced approach named StarGAN[20] was proposed in [21]. Before the StarGAN approach, many studies had shown promising results in image-to-image translations for two domains in general. However, few studies had focused on handling more than two domains, as different models should be built independently for every pair of image domains. StarGAN proposed a novel approach able to perform image-to-image translations for multiple domains using only a single model. The authors trained a conditional attribute transfer network via attribute classification loss and cycle consistency loss. Good visual results were achieved compared with previous approaches. However, it sometimes includes undesired

[17] https://github.com/shaoanlu/faceswap-GAN

[18] https://github.com/Guim3/IcGAN

[19] https://github.com/facebookresearch/FaderNetworks

[20] https://github.com/yunjey/stargan/blob/master/README.md

TABLE V
FACIAL ATTRIBUTES: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Bharati *et al.* (2016) [77]	Face Patches	RBM	Overall Acc. = 96.2% Overall Acc. = 87.1%	Own (Celebrity Retouching, ND-IIITD Retouching)

Tariq *et al.* (2018)
[78]
	Image-related	CNN	AUC = 99.9%
			AUC = 74.9%
				Own
				(ProGAN,
				Adobe Photoshop)

Wang *et al.* (2019) [30]	CNN Neuron Behavior	SVM	Acc. = 84.7%	Own (InterFaceGAN/StyleGAN)
Jain *et al.* (2019) [79]	Face Patches	CNN + SVM	Overall Acc. = 99.6% Overall Acc. = 99.7%	Own (ND-IIITD Retouching, StarGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.9%** **EER = 1.0%**	**DFFD (FaceApp/StarGAN)**
Wang *et al.* (2019) [80]	Image-related	DRN	AP = 99.8%	Own (Adobe Photoshop)
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	Acc. = 99.4%	Own (StarGAN/CycleGAN)
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (Glow/StarGAN)
Zhang *et al.* (2019) [81]	Frequency Domain	GAN Discriminator	Acc. = 100%	Own (StarGAN/CycleGAN)

modifications from the input face image such as the colour of the skin.

Almost at the same time He *et al.* proposed in [91] attGAN[21], a novel approach that removes the strict attribute-independent constraint from the latent representation, and just applies the attribute-classification constraint to the generated image to guarantee the correct change of the attributes. AttGAN provides state-of-the-art results on realistic attribute editing with other facial details well preserved.

One of the latest approaches proposed in the literature is STGAN[22] [92]. As can be seen, attribute editing generally can be tackled by incorporating encoder-decoders and GANs. However, as commented Liu *et al.* [92], the bottleneck layer in the encoder-decoder usually provides blurry and low quality editing results. To improve this, the authors presented and incorporated selective transfer units with a encoder-decoder for simultaneously improving the attribute manipulation ability and the image quality. As a result, STGAN has recently outperformed the state-of-the-art in facial attribute editing.

Despite of the fact that the code of most facial attribute manipulation approaches are publicly available, the lack of public databases and experimental protocols results crucial when comparing among different manipulation detection approaches, as it is not possible to perform a fair comparison among studies. Up to now, to the best of our knowledge, the DFFD database [7] seems to be the only public database that considers this type of facial manipulations. This database comprises 18,416 and 79,960 fake images generated through FaceApp and StarGAN approaches, respectively.

B. Manipulation Detection

Facial attribute manipulations have been originally studied in the field of face recognition in order to see how robust biometric systems are against physical factors such as plastic surgery, cosmetics, makeup or occlusions [94]–[98]. However, it has not been until the recent success of mobile applications such as FaceApp, that the research community has been motivated to detect digital face attribute manipulations. Table V provides a comparison of the most relevant approaches in this area. We include for each study information related to the features, classifiers, best performance, and databases for research.

In [77], Bharati *et al.* proposed a deep learning approach based on a Restricted Boltzmann Machine (RBM) in order to detect digital retouching of face images. The input of the detection system consisted of face patches in order to learn discriminative features to classify each image as original or retouched. Regarding the databases, the authors generated two fake databases from the original ND-IIITD database (collection B [99]) and a set of celebrity facial images downloaded from the Internet. Fake images were generated using the professional software PortraitPro Studio Max[23], considering aspects such as skin texture, shape of eyes, nose, lips and overall face, prominence of smile, lip shape, and eye colour. Their proposed approach achieved overall accuracies for manipulation detection of 96.2% and 87.1% for the celebrity and ND-IIITD retouching databases, respectively.

Tariq *et al.* proposed in [78] the use of CNNs in order to detect facial attribute manipulations. Different CNN architectures such as VGG16 [37], VGG19 [37], ResNet [70], or XceptionNet [74], among others, were evaluated. For the real face images, the CelebA database [23] was considered.

[21] https://github.com/LynnHo/AttGAN-Tensorflow
[22] https://github.com/csmliu/STGAN
[23] https://www.anthropics.com/portraitpro/

Regarding the fake images, two different approaches were considered: *i)* machine approaches based on GANs, in particular ProGAN [26], and *ii)* manual approach based on Adobe Photoshop CS6, including manipulations such as makeup, glasses, sunglasses, hair, and hats. For the experimental evaluation, different sizes of the images were considered (from 32×32 to 256×256 pixels). A final 99.99% AUC was obtained for the machine-created scenario whereas for the human-created scenario this value decreased to a final 74.9% AUC for the best CNN model. Thus, a high degradation of the system performance is observed among machine- and human-created fake images.

In [30], Wang *et al.* conjectured that monitoring neuron behavior could also serve as an asset in detecting fake faces since layer-by-layer neuron activation patterns may capture more subtle features that are important for the facial manipulation detection system. Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (VGG-Face [37], OpenFace [38], and FaceNet [39]), and then trained a SVM for the final classification. The authors tested their proposed approach using real faces from CelebA-HQ [26] and FFHQ [19] databases and synthetic faces created through InterFaceGAN [40] and StyleGAN [19], achieving for the best performance a final 84.7% accuracy using the FaceNet model.

Jain *et al.* proposed in [79] a detection system based on a CNN architecture composed of 6 convolutional layers and 2 fully-connected layers. Additionally, residual connections were considered inspired by a ResNet architecture [70]. Similar to [77], the input of the CNN system consists of non-overlapping face patches to learn discriminative features. Finally, a SVM was used for the final classification. Regarding the experimental framework, the ND-IIITD retouched database presented in [77] was considered. Additionally, the authors considered fake images created through the StarGAN approach [21], trained using the CelebA database [23]. In general, good detection results were achieved in both manipulation approaches, achieving almost 100% accuracy.

Stehouwer *et al.* carried out in [7] a complete analysis of different facial manipulation methods. They proposed to use attention mechanisms to process and improve the feature maps of CNN models. Regarding the facial attribute manipulations, two different approaches were considered: *i)* fake images created through the public FaceApp software, with up to 28 different available filters considering aspects such as hair, age, glasses, beard, and skin colour, among others; and *ii)* fake images created through the StarGAN approach [21], with up to 40 different filters. Their proposed approach was tested using their novel database DFFD, achieving very good results close to 1.0% EER (and 99.9% of AUC).

Wang *et al.* carried out in [80] an interesting research using publicly available commercial software from Adobe Photoshop (Face-Aware Liquify tool [100]) in order to synthesise new faces, and also a professional artist in order to manipulate 50 real photographs. The authors began running a human study through Amazon Mechanical Turk (AMT), showing real and fake images to the participants and asking them to classify each image into one of the classes. The results achieved remark how challenging the task is for humans, with a final 53.5% of accuracy, close to chance (50%). After the human study, the authors proposed an automatic detection system based on Deep Recurrent Networks, achieving manipulation detection performances of 99.8% and 97.4% for automatic and manual face synthesis manipulation.

As described in Sec. III-B for face synthesis, Nataraj *et al.* proposed in [31] a detection system inspired by steganalysis and natural image statistics. The authors created a new fake dataset based on facial attribute manipulations using the StarGAN approach [21] trained through the CelebA database [23]. Their proposed detection approach achieved a final 99.4% accuracy for the best result.

The work [32] by Marra *et al.* also described in Sec. III-B was able to correctly perform discrimination when new GANs were presented to the network and achieved a 99.3% accuracy for their proposed detection approach, based on the XceptionNet model.

Finally, Zhang *et al.* proposed in [81] a detection system based on the spectrum domain, rather than the raw image pixels. Given an image as input, they applied a 2D DFT to each of the RGB channels, getting one frequency image per channel. Regarding the classifier, they proposed AutoGAN, which is a GAN simulator that can synthesise GAN artifacts in any image without needing to access any pre-trained GAN model. The generalisation capacity of their proposed approach was tested using unseen GAN models. In particular, StarGAN [21] and GauGAN [82] were considered in the evaluation. For the StarGAN approach, good detection results were achieved using the frequency domain (100%). However, for the GauGAN approach, a high degradation of the system performance, 50% accuracy, was observed. The authors claimed that this was produced due to the generator of the GauGAN is drastically different from the CycleGAN (used in training).

VI. Facial Expression

A. Manipulation Techniques and Public Databases

To the best of our knowledge, the only available database to date focused on facial expression manipulation is FaceForensics++ [6], an extension of FaceForensics [61]. Initially, FaceForensics database was focused only on the Face2Face approach [22]. This is a computer graphics approach that transfers the expression of a source video to a target video while maintaining the identity of the target person. This was carried out through manual keyframe selection. Concretely, the first frames of each video were used to obtain a temporary face identity (i.e., a 3D model), and track the expression over the remaining frames. Then, fake videos were generated by transferring the source expression parameters of each frame (i.e., 76 Blendshape coefficients) to the target video. Later on, the same authors presented in FaceForensics++ a new learning approach based on NeuralTextures [102]. This is a rendering approach that uses the original video data to learn a neural texture of the target person, including a rendering network. In particular, the authors considered in their implementation a patch-based GAN-loss as used in Pix2Pix [82]. Only the facial expression

TABLE VI
FACIAL EXPRESSION: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Afchar *et al.* (2018) [53]	Mesoscopic Level	CNN	Acc. = 83.2%	FF++ (Face2Face, LQ)
			Acc. = 93.4%	FF++ (Face2Face, HQ)
			Acc. = 96.8%	FF++ (Face2Face, RAW)
			Acc. $\simeq$ 75%	FF++ (NeuralTextures, LQ)
			Acc. $\simeq$ 85%	FF++ (NeuralTextures, HQ)
			Acc. $\simeq$ 95%	FF++ (NeuralTextures, RAW)
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. $\simeq$ 91%	FF++ (Face2Face, LQ)
			Acc. $\simeq$ 98%	**FF++ (Face2Face, HQ)**
			Acc. $\simeq$ 100%	**FF++ (Face2Face, RAW)**
			Acc. $\simeq$ 81%	FF++ (NeuralTextures, LQ)
			Acc. $\simeq$ 93%	**FF++ (NeuralTextures, HQ)**
			Acc. $\simeq$ 99%	**FF++ (NeuralTextures, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression, MLP	AUC = 86.6%	FF++ (Face2Face, RAW)
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	EER = 7.1%	FF++ (Face2Face, HQ)
			EER = 7.8%	FF++ (NeuralTextures, HQ)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.4%** **EER = 3.4%**	**FF++ (Face2Face, -)**
Amerini *et al.* (2019) [101]	Inter-Frame Dissimilarities	CNN + Optical Flow	Acc. = 81.6%	FF++ (Face2Face, -)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**Acc. = 94.3**	**FF++ (Face2Face, LQ)**

corresponding to the mouth was modified. It is important to remark that all data is available on the FaceForensics++ GitHub[24]. In total, there are 1000 real videos extracted from Youtube. Regarding the manipulated videos, 2000 fake videos are available (1000 videos for each considered approach). In addition, it is important to highlight that different video quality levels are considered, in particular: *i)* RAW (original quality), *ii)* HQ (constant rate quantization parameter equal to 23), and *iii)* LQ (constant rate quantization parameter equal to 40). This aspect simulates the video processing techniques usually applied in social networks.

Although both Face2Face and NeuralTextures have been traditionally considered as facial expression manipulation techniques, many other approaches allow to modify the facial expression nowadays. For example, mobile applications such as FaceApp[25] allows to easily change the level of smiling, from happier to angrier. These approaches are based on current GAN architectures. For example, Choi *et al.* showed in [21] the potential of StarGAN to change an input image to different expression levels such as angry, happy, neutral, sad, surprised, and fearful. Other recent GAN approaches that improve both the image quality of the fake images and the control editing of the parameters are InterFaceGAN [40], UGAN [103], STGAN [92], and AttGAN [91].

B. Manipulation Detection

One of the articles that fostered the development of novel detection techniques in facial expression manipulation is [16]. In that study, the authors showed how technology was able to synthesise high-quality videos of a person (Obama in this case) changing what he is really saying in a target video [104]. The impressive results achieved motivated the research community to develop robust detection techniques. Table VI provides a comparison of the most relevant approaches in this area. For each study we include information related to the features, classifiers, best performance, and databases. We highlight in **bold** the best results achieved for the only public database, FaceForensics++. It is important to remark that in some cases, different evaluation metrics are considered (e.g., AUC and EER), which makes it difficult to perform a fair comparison among the studies.

The following methods were discussed in Sect. IV-B for face swap detection. Here we summarise the results achieved by them in detecting facial expression manipulation. In [53], the proposed approach was tested using the Face2Face fake videos from the FaceForensics++ database [6], achieving in general good results, especially for RAW-quality videos. The same approach was later on tested in [6] against NeuralTextures fake videos, obtaining lower accuracy results compared with the Face2Face scenario. In [61] by Rössler *et al.*, the detection system based on XceptionNet provided the best results in both Face2Face and NeuralTextures manipulations, close to 100% on RAW quality. In addition, the detection systems were evaluated considering different video quality levels in order to simulate the video processing of many social networks. In this real scenario, the accuracy of all detection systems were degraded with the video quality, as it happens for face swap manipulations. In [57] by Matern *et al.*, the proposed approach was tested using the FaceForensics++ database, but only the Face2Face manipulation technique, achieving a final 86.6% AUC for the best performance. In [58] the proposed approach was evaluated with the FaceForensics++ database. For the Face2Face method, they achieved a 7.1% EER on

[24] https://github.com/ondyari/FaceForensics

[25] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341

HQ videos whereas for the NeuralTexture method, the EER increased a bit more to a final 7.8% EER. In [7] by Stehouwer *et al.*, the proposed detection approach was tested using the DFFD database, which for the facial expression manipulation scenario is based only on data from FaceForensics++ database. The proposed approach achieved an AUC = 99.4% and EER = 3.4%. In [60] by Sabir *et al.*, the proposed detection approach was tested with FaceForensics++ database, achieving AUC results of 94.3% for the Face2Face technique. Only the low-quality videos were considered in the analysis. Finally, in [101], Amerini *et al.* proposed the adoption of optical flow fields to exploit possible inter-frame dissimilarities. The optical flow is a vector field computed among two consecutive frames to extract apparent motion between the observer and the scene itself. The use of this approach is motivated as fake videos should be more appreciable in the optical flow matrices due to the unusual movement of lips, eyes, etc. Preliminary results were obtained using both VGG16 and ResNet50 networks, obtaining an Acc. = 81.6% for the best performance.

VII. Concluding Remarks

Motivated by the ongoing success of digital face manipulations, specially DeepFakes, this survey provides an exhaustive panorama of the field, including details of up-to-date: *i)* types of facial manipulations, *ii)* facial manipulation techniques, *iii)* public databases for research, and *iv)* benchmarks for the detection of each facial manipulation group, including key results achieved by the most representative manipulation detection approaches.

Generally speaking, and despite the significant improvement in face manipulation achieved recently, most face manipulation examples included in popular databases are now easily detected, not only by humans but also by machines. This fact has been demonstrated in most of the benchmarks included in this survey, achieving very low error rates by modern face manipulation detection systems. However, a number of challenges still remain in this field.

So far today, most approaches for fake detection are focused on controlled scenarios, e.g., training and testing detection systems considering the same image compression level. However, this approach seems not to be the most appropriate one for real scenarios. New approaches need to be developed in order to overcome such variations of the image/video (compression level, noise, blur, etc.). This aspect has been already remarked in different studies [6], [11], [29], observing a high degradation of the fake detection performance when considering in-the-wild scenarios. Another important aspect needs to be covered: how robust are detection systems to unseen face manipulation attacks not considered in training? For example, future GAN approaches. In general, a poor generalisation capacity is observed in most cases even for state-of-the-art detection systems [17], [32].

Finally, at image level, face manipulations such as the entire face or facial attributes have already achieved very realistic results thanks to novel GAN approaches such as StyleGAN [19]. However, most systems find the face manipulation detection task easy to perform due to the GAN “fingerprints” included in the fake images. What if we are able to remove those fingerprints? A preliminary study of this idea was carried out in [11] through the use of autoencoders and image quality degradation, observing a considerable degradation of the detection systems.

All these aspects, together with the improvement and new development of future GAN approaches and the recent DeepFake Detection Challenge (DFDC) organised by Facebook and others will foster the new generation of realistic fake images/videos [105] together with more advanced techniques for face manipulation detection.

Acknowledgments

This work has been supported by projects: BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), Bio-Guard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017).

References

[1] P. Korshunov and S. Marcel, “Deepfakes: a New Threat to Face Recognition? Assessment and Detection,” *arXiv preprint arXiv:1812.08685*, 2018.

[2] D. Citron, “How DeepFake Undermine Truth and Threaten Democracy,” 2019. [Online]. Available: https://www.ted.com

[3] R. Cellan-Jones, “Deepfake Videos Double in Nine Months,” 2019. [Online]. Available: https://www.bbc.com/news/technology-49961089

[4] BBC Bitesize, “Deepfakes: What Are They and Why Would I Make One?” 2019. [Online]. Available: https://www.bbc.co.uk/bitesize/articles/zfkwcqt

[5] M. Stamm and K. Liu, “Forensic Detection of Image Manipulation Using Statistical Intrinsic Fingerprints,” *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[6] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in *Proc. International Conference on Computer Vision*, 2019.

[7] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain, “On the Detection of Digital Face Manipulation,” *arXiv preprint arXiv:1910.01717*, 2019.

[8] J. Galbally, S. Marcel, and J. Fierrez, “Biometric Anti-Spoofing Methods: A Survey in Face Recognition,” *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[9] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, “Biometrics Systems Under Spoofing Attack: an Evaluation Methodology and Lessons Learned,” *IEEE Signal Processing Magazine*, 2015.

[10] S. Marcel, M. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing (2nd Edition)*, 2019.

[11] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, and H. Proença, “Real or Fake? Spoofing State-Of-The-Art Face Synthesis Detection Systems,” *arXiv preprint arXiv:1911.05351*, 2019.

[12] C. Canton, L. Davis, E. Delp, P. Flynn, S. McCloskey, L. Leal-Taixe, P. Natsev, and C. Bregler, “Applications of Computer Vision and Pattern Recognition to Media Forensics,” in *Conference on Computer Vision and Pattern Recognition*, 2019. [Online]. Available: https://sites.google.com/view/mediaforensics2019

[13] B. Biggio, P. Korshunov, T. Mensink, G. Patrini, D. Rao, and A. Sadhu, “Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes,” in *International Conference on Machine Learning*, 2019. [Online]. Available: https://sites.google.com/view/audiovisualfakes-icml2019/

[14] L. Verdoliva and P. Bestagini, “Multimedia Forensics,” in *ACM Multimedia*, 2019. [Online]. Available: https://acmmm.org/tutorials/#tut3

[15] C. Bregler, M. Covell, and M. Slaney, “Video Rewrite: Driving Visual Speech with Audio,” *Computer Graphics*, vol. 31, no. 2, pp. 353–361, 1997.

[16] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning Lip Sync From Audio,” *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.

[17] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A New Dataset for DeepFake Forensics,” *arXiv preprint arXiv:1909.12962*, 2019.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. Advances in Neural Information Processing Systems*, 2014.

[19] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[20] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, 2018.

[21] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[22] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-Time Face Capture and Reenactment of RGB Videos," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. International Conference on Computer Vision*, 2015.

[24] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning Face Representation From Scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[25] Q. Cao, , L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces Across Pose and Age," in *Proc. International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 67–74.

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *Proc. International Conference on Learning Representations*, 2018.

[27] 100,000 Faces Generated by AI, 2018. [Online]. Available: https://generated.photos/

[28] S. McCloskey and M. Albright, "Detecting GAN-Generated Imagery Using Color Cues," *arXiv preprint arXiv:1812.08247*, 2018.

[29] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images," in *Proc. International Conference on Computer Vision*, 2019.

[30] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, "FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces," *arXiv preprint arXiv:1909.06122*, 2019.

[31] L. Nataraj, T. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. Bappy, and A. Roy-Chowdhury, "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," *arXiv preprint arXiv:1903.06836*, 2019.

[32] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental Learning for the Detection and Classification of GAN-Generated Images," in *Proc. International Workshop on Information Forensics and Security*, 2019.

[33] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019, pp. 63–72.

[34] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. International Conference on Learning Representations*, 2018.

[35] M. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The Cramer Distance as a Solution to Biased Wasserstein Gradients," *arXiv preprint arXiv:1705.10743*, 2017.

[36] M. Binkowski, D. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. International Conference on Learning Representations*, 2018.

[37] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proc. British Machine Vision Conference*, 2015.

[38] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A General-Purpose Face Recognition Library with Mobile Applications," in *CMU School of Computer Science*, 2016.

[39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.

[40] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the Latent Space of GANs for Semantic Face Editing," *arXiv preprint arXiv:1907.10786*, 2019.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[42] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. International Conference on Computer Vision*, 2017.

[43] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[44] D. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Proc. Advances in Neural Information Processing Systems*, 2018.

[45] M. Huh, A. Liu, A. Owens, and A. Efros, "Fighting Fake News: Image Splice Detection Via Learned Self-Consistency," in *Proc. of the European Conference on Computer Vision*, 2018.

[46] P. Zhou, X. Han, V. Morariu, and L. Davis, "Learning rich features for image manipulation detection," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[47] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in *Proc. International Workshop on Information Forensics and Security*, 2018.

[48] C. Sanderson and B. Lovell, "Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference," in *Proc. International Conference on Biometrics*, 2009.

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[50] Google AI, "Contributing Data to Deepfake Detection Research," 2019. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[51] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[52] P. Zhou, X. Han, V. Morariu, and L. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[53] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *Proc. International Workshop on Information Forensics and Security*, 2018.

[54] D. Güera and E. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. International Conference on Advanced Video and Signal Based Surveillance*, 2018.

[55] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019.

[56] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[57] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019.

[58] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos," *arXiv preprint arXiv:1906.06876*, 2019.

[59] S. Agarwal and H. Farid, "Protecting World Leaders Against Deep Fakes," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[60] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[61] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," *arXiv preprint arXiv:1803.09179*, 2018.

[62] P. Pérez, M. Gangnet, and A. Blake, "Poisson Image Editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.

[64] P. Korshunov and S. Marcel, "Speaker Inconsistency Detection in Tampered Video," in *Proc. European Signal Processing Conference*, 2018.

[65] J. Galbally, S. Marcel, and J. Fierrez, "Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint and Face

Recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 710–724, 2014.

[66] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[67] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2009.

[68] D. King, "DLib-ML: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[69] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[71] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[72] B. Bayar and M. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2016.

[73] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks," in *Proc. Workshop on Information Forensics and Security*, 2017.

[74] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[75] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 2016.

[76] T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *Proc. International Conference on Automatic Face & Gesture Recognition*, 2018.

[77] A. Bharati, R. Singh, M. Vatsa, and K. Bowyer, "Detecting Facial Retouching Using Supervised Deep Learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903–1913, 2016.

[78] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo, "Detecting Both Machine and Human Created Fake Face Images in the Wild," in *Proc. International Workshop on Multimedia Privacy and Security*, 2018, pp. 81–87.

[79] A. Jain, R. Singh, and M. Vatsa, "On Detecting GANs and Retouching based Synthetic Alterations," in *Proc. International Conference on Biometrics Theory, Applications and Systems*, 2018.

[80] S. Wang, O. Wang, A. Owens, R. Zhang, and A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop," *arXiv preprint arXiv:1906.05856*, 2019.

[81] X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," *arXiv preprint arXiv:1907.06515*, 2019.

[82] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[83] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. Efros, O. Wang, and E. Shechtman, "Toward Multimodal Image-to-Image Translation," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[84] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks," in *Proc. International Conference on Machine Learning*, 2017.

[85] D. Bau, J. Zhu, H. Strobelt, B. Zhou, J. Tenenbaum, W. Freeman, and A. Torralba, "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks," *arXiv preprint arXiv:1811.10597*, 2018.

[86] G. Perarnau, J. V. D. Weijer, B. Raducanu, and J. Álvarez, "Invertible Conditional GANs for Image Editing," in *Proc. Advances in Neural Information Processing Systems Workshops*, 2016.

[87] M. Li, W. Zuo, and D. Zhang, "Deep Identity-Aware Transfer of Facial Attributes," *arXiv preprint arXiv:1610.05586*, 2016.

[88] W. Shen and R. Liu, "Learning Residual Images for Face Attribute Manipulation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[89] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader Networks: Manipulating Images by Sliding Attributes," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[90] T. Xiao, J. Hong, and J. Ma, "ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes," in *Proc. European Conference on Computer Vision*, 2018.

[91] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing*, 2019.

[92] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[93] M. Mirza and S. S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.

[94] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 12, pp. 1977–1981, 2005.

[95] A. Dantcheva, C. Chen, and A. Ross, "Can Facial Cosmetics Affect the Matching Accuracy of Face Recognition Systems?" in *Proc. International Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 391–398.

[96] N. Kose, L. Apvrille, and J. Dugelay, "Facial Makeup Detection Technique based on Texture and Shape Analysis," in *Proc. International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[97] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa, "Evading Face Recognition via Partial Tampering of Faces," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[98] C. Rathgeb, A. Dantcheva, and C. Busch, "Impact and Detection of Facial Beautification in Face Recognition: An Overview," *IEEE Access*, vol. 7, pp. 152 667–152 678, 2019.

[99] P. Flynn, K. Bowyer, and P. Phillips, "Assessment of Time Dependency in Face Recognition: An Initial Study," in *Proc. International Conference on Audio-and Video-Based Biometric Person Authentication*, 2003.

[100] Adjust and exaggerate facial features. Adobe Photoshop, 2016. [Online]. Available: https://helpx.adobe.com/photoshop/how-to/face-aware-liquify.html

[101] I. Amerini, L. Galteri, R. Caldelli, and A. Bimbo, "Deepfake Video Detection through Optical Flow based CNN," in *Proc. International Conference on Computer Vision*, 2019.

[102] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred Neural Rendering: Image Synthesis using Neural Textures," *ACM Transactions on Graphics*, vol. 38, no. 66, pp. 1–12, 2019.

[103] D. Zhu, S. Liu, W. Jiang, C. Gao, T. Wu, and G. Guo, "UGAN: Untraceable GAN for Multi-Domain Face Translation," *arXiv preprint arXiv:1907.11418*, 2019.

[104] S. Suwajanakorn, "Fake Videos of Real People – and How to Spot Them," 2019. [Online]. Available: https://www.ted.com

[105] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," *arXiv preprint arXiv:1912.04958*, 2019.

Ruben Tolosana received the M.Sc. degree in Telecommunication Engineering, and his Ph.D. degree in Computer and Telecommunication Engineering, from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In April 2014, he joined the Biometrics and Data Pattern Analytics - BiDA Lab at the Universidad Autonoma de Madrid, where he is currently collaborating as a PostDoctoral researcher. Since then, Ruben has been granted with several awards such as the FPU research fellowship from Spanish MECD (2015), and the European Biometrics Industry Award (2018). His research interests are mainly focused on signal and image processing, pattern recognition, and machine learning, particularly in the areas of face manipulation, human-computer interaction and biometrics. He is author of several publications and also collaborates as a reviewer in many different high-impact conferences (e.g., ICDAR, IJCB, ICB, BTAS, EUSIPCO, etc.) and journals (e.g., IEEE TPAMI, TCYB, TIFS, TIP, ACM CSUR, etc.). Finally, he has participated in several National and European projects focused on the deployment of biometric security through the world.

Ruben Vera-Rodriguez received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Spain, in 2006, and the Ph.D. degree in electrical and electronic engineering from Swansea University, U.K., in 2010. Since 2010, he has been affiliated with the Biometric Recognition Group, Universidad Autonoma de Madrid, Spain, where he is currently an Associate Professor since 2018. His research interests include signal and image processing, pattern recognition, and biometrics, with emphasis on signature, face, gait verification and forensic applications of biometrics. He is actively involved in several National and European projects focused on biometrics. Ruben has been Program Chair for the IEEE 51st International Carnahan Conference on Security and Technology (ICCST) in 2017; and the 23rd Iberoamerican Congress on Pattern Recognition (CIARP 2018) in 2018.

Julian Fierrez received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002, he has been with the Biometric Recognition Group, Universidad Politecnica de Madrid. Since 2004, he has been with the Universidad Autonoma de Madrid, where he is currently an Associate Professor. From 2007 to 2009, he was a Visiting Researcher with Michigan State University, USA, under a Marie Curie Fellowship. His research interests include signal and image processing, pattern recognition, and biometrics, with an emphasis on multibiometrics, biometric evaluation, system security, forensics, and mobile applications of biometrics. He has been actively involved in multiple EU projects focused on biometrics (e.g., TABULA RASA and BEAT), and has attracted notable impact for his research. He was a recipient of a number of distinctions, including the EAB European Biometric Industry Award 2006, the EURASIP Best Ph.D. Award 2012, the Miguel Catalan Award to the Best Researcher under 40 in the Community of Madrid in the general area of science and technology, and the 2017 IAPR Young Biometrics Investigator Award. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Aythami Morales received the M.Sc. degree in telecommunication engineering from the Universidad de Las Palmas de Gran Canaria in 2006 and the Ph.D. degree from La Universidad de Las Palmas de Gran Canaria in 2011. Since 2017, he is Associate Professor with the Universidad Autonoma de Madrid. He has conducted research stays at the Biometric Research Laboratory, Michigan State University, the Biometric Research Center, Hong Kong Polytechnic University, the Biometric System Laboratory, University of Bologna, and the Schepens Eye Research Institute. He has authored over 70 scientific articles published in international journals and conferences. He has participated in national and EU projects in collaboration with other universities and private entities, such as UAM, UPM, EUPMt, Indra, Union Fenosa, Soluziona, or Accenture. His research interests are focused on pattern recognition, computer vision, machine learning, and biometrics signal processing. He has received awards from the ULPGC, La Caja de Canarias, SPEGC, and COIT.

Javier Ortega-Garcia received the M.Sc. degree in electrical engineering and the Ph.D. degree (cum laude) in electrical engineering from Universidad Politecnica de Madrid, Spain, in 1989 and 1996, respectively. He is currently a Full Professor at the Signal Processing Chair in Universidad Autonoma de Madrid - Spain, where he holds courses on biometric recognition and digital signal processing. He is a founder and Director of the BiDA-Lab, Biometrics and Data Pattern Analytics Group. He has authored over 300 international contributions, including book chapters, refereed journal, and conference papers. His research interests are focused on biometric pattern recognition (on-line signature verification, speaker recognition, human-device interaction) for security, e-health and user profiling applications. He chaired Odyssey-04, The Speaker Recognition Workshop, ICB-2013, the 6th IAPR International Conference on Biometrics, and ICCST2017, the 51st IEEE International Carnahan Conference on Security Technology.